# Evaluation of Metadata Change in Authority Data over Time: an Effect of a Standard Evolution

**Oksana L. Zavalina**
*University of North Texas, USA.*
Oksana.Zavalina@unt.edu

**Vyacheslav Zavalin**
*University of North Texas, USA.*
VyacheslavZavalin@my.unt.edu

## ABSTRACT

Information community creates, maintains and shares authority data through large-scale databases of standardized digital records that describe persons, institutions, places, events, and works, as well as relations between them. This submission presents some results of the content analysis study that explores the authority data change over time in response to change in standards. We analysed over 400 thousand of authority records that comply to the new standard, Resource Description and Access (RDA) and are available through the OCLC database. Records were obtained at two data collection points, with an interval of 22 months. Our analysis identified RDA-based authority data elements that are widely applied and the ones that need more attention by record creators. Findings reveal a significant increase over time in the level of application of some data elements, including several of the Linked Data-enabling elements. This study contributes to the understanding of metadata change and its relation to functionality of authority records and improved information access. Directions for future research are suggested.

## KEYWORDS

Metadata analysis, authority data, Linked Data, MARC, RDA.

## INTRODUCTION

The content of metadata records is influenced by various environmental change over time, including changes in national and international standards for record creation, etc. (Thornburg and Oskins, 2007). To keep pace with these environmental changes, metadata records to be regularly reviewed and updated. This applies both to bibliographic metadata that represents information objects, and to authority data that has long been used to facilitate access to information. Authority records represent in a standard way the names of persons, organizations, meetings, places and works related to information objects (e.g., as composers, translators, etc.) and subjects covered in these information objects.

The process of authority records creation is guided by data content standards that instruct what information to record and how to represent it, and data encoding standards for technical formatting and exchange of authority data. Since 1960s, authority data have been created, maintained and exchanged in digital form through the large-scale databases that rely on the international data encoding standard Machine-Readable Cataloging (MARC) Format for Authority Data (http://www.loc.gov/marc/authority/). The data in MARC authority records are organized into fixed and variable fields. Most of the variable fields (identified by a three-character numeric tag, e.g., 100), include multiple data elements (e.g., #a, #2, etc.). The authority data are freely available for harvesting individually and as record sets in their native MARC format, as well as (in the recent years) in Linked Data form in serializations of Resource Description Framework (RDF).

Both data content and data encoding standards evolve over time to reflect the changes in information environment and to meet the emerging requirements (e.g., the requirements of Semantic Web). In 2013, the new Resource Description and Access (RDA) standard designed to better support Linked Data officially replaced its predecessor as a data content standard for authority data. Since the early stages of RDA development, the active work is ongoing on aligning MARC with RDA. In the process of alignment, numerous new MARC record fields and subfields have been added, with the most recent revisions of RDA-based MARC Authority Standard – update no. 25 – completed in December 2017. For example, the "heading information" block of MARC authority fields has been greatly expanded through addition of 20 new 3XX MARC fields (e.g., 373 Associated Group).

Very few published papers so far discussed RDA implementation in authority records or evaluated implementation results. Niu (2013) and Southwick (2015) discussed the process of RDA-based MARC authority records creation or conversion of existing records into Linked Data form. Kimura (2015) compared data elements in over a million of personal, corporate, and meeting name authority records created in China, Japan, and Korea with data elements defined in RDA standard. Thompson's study (Thompson, 2016) focused on the application of the RDA-based field 375 Gender in a small sample of personal name authority records. Moulaison (2015) assessed the use of 7 RDA-based data elements in personal name authority records in an academic library consortia authority file one year after official adoption of RDA. Zavalina and Zavalin (2017) took a broader approach

and examined levels of application of 35 RDA-based MARC fields and their subfields in 5 types of authority records 3 years after RDA adoption.

Despite the importance of measuring metadata change that was pointed out for information quality (Stvilia et al., 2004; Stvilia & Gasser, 2008), little research on metadata change has been published; all of the available studies focus on change in descriptive metadata, with no investigation of authority data change. The overall scarcity of metadata change studies is in part due to the shortage of information systems that provide a metadata versioning functionality. Several quantitative studies attempted to identify and measure change in metadata records in digital libraries that enable metadata versioning (e.g., Tarver, Zavalina & Phillips, 2016; Zavalina, Phillips & Tarver, 2017). A qualitative research project (e.g., Zavalina et al., 2015, 2016; Zavalina, Shakeri, & Kizhakkethil, 2015; Zavalina, Shakeri, Kizhakkethil, & Phillips, 2018) categorized metadata change in digital library metadata and in traditional library metadata.

The transition to RDA standard significantly impacts authority data that are crucial for providing adequate access to information. A research gap in the area of implementation of RDA standard in authority data exists. Little is known as to how the implementation of many new, RDA-based, MARC data elements is reflected in the authority data that is currently in use, and how the RDA-based authority records change over time.

## METHODS

This study addresses the gap through quantitative empirical evaluation of the large representative dataset. Investigation is guided by the following research questions: What is the level of application of the new RDA-based data elements of MARC authority records and how does this level change over time? How are the Linked Data enabling elements of RDA applied in the existing authority data and how this changes over time?

MARC Edit (http://marcedint.reesenet.net) tool – Raw (ADV) search in Z39.50/SRU Client – was used for collecting the data to answer these research questions. The data was collected from the United States Library of Congress Authorities (https://authorities.loc.gov/) database, available for harvesting through OCLC, at two points in time, with an interval of 22 months: in early 2016 and in late 2017. A total of 408551 RDA-based authority records were retrieved in 2016. The same records, based on the unique records IDs, were collected in 2017. The MARC Edit Z39.50/SRU Client search in 2017 resulted in retrieving all but 26 records from the 2016 dataset; the 26 missing records had likely been deleted from the OCLC authority database between the two data collection points. Preliminary analysis identified, based on evaluation of data in the field 005 Date and Time of Latest Transaction, 35472 records that underwent changes between the two data collection points. This set included five types of authority records: corporate, geographic, meeting and personal names, as well as uniform titles. MARC Edit, Microsoft Excel and Sublime text editor were used in processing and analyzing the data.

In the next section, we present some results of the comparative content analysis of 2016 and 2017 versions of the changed records. We measured the utilization and assessed the level of change in application of RDA-based MARC authority variable fields (with their respective subfields) and RDA-based subfields of non-RDA-specific variable fields (#i Relationship Information, #4 Relationship #u Uniform Resource Identifier, #1 Real-World Object URI, #2 Source of Data, #0 Authority Record Number).

## FINDINGS AND DISCUSSION

Each record in the dataset underwent editing between the two data collection points. Less than half of 35472 records had been edited prior to initial data collection, as indicated by a total of 13786 occurrences of MARC 21 field 040 subfield #d an instance of each is automatically added upon every editing event. It is worth noting that each editing event could include multiple changes to one or more fields of the record. In the 2017 versions of authority records, the number of occurrences of this subfield increased to 49687, meaning one editing event per record on average over the period of 22 months. This is lower than the level of editing activity observed by previous research for bibliographic metadata during the first years after transition to RDA.

Our findings (Table 1) demonstrate that almost all of the RDA-based authority records that underwent changes between two data collection points were meeting name and personal name records (99.76% collectively: 60.15% and 39.6% respectively). The proportion of corporate name, geographic name and uniform title records among the edited records was minimal, between 0.02% and 0.17%. While in the total dataset of 408551 RDA-based records collected in 2016, meeting and personal names also represented the majority (86.39% collectively: 60.5% and 26.39%), the disproportional distribution of edited metadata records is an indication of a higher attention of metadata editors towards meetings and persons. This record-type-level finding is supported by data-element-level findings (Figures 1 and 2).

| | meeting name | personal name | corporate name | uniform title | geographic name |
|---|---|---|---|---|---|
| % in total dataset (408551) | 60.50% | 26.39% | 5.03% | 4.45% | 3.63% |
| % in edited records (35472) | 60.16% | 39.60% | 0.17% | 0.04% | 0.02% |

**Table 1. Distribution by authority record type**

A total of 17 RDA-based MARC fields were observed in the records that underwent changes between two data collection points. This represents 50% of all RDA-based fields in MARC authority standard at the time of 2017 data collection. None of the records in our dataset was found to include RDA-based fields 075, 147, 162, 336, 348, 385, 386, 388, 447, 462, 547, 562, 677, 700, 730, 747 or 762. In 2016, the most frequently occurring (in over 99% of records) RDA-based fields were 046 Special Coded Dates, 377 Associated Language, and 370 Associated Place. In 2017, the level of application of two more fields significantly increased and reached 99% or more of records: 373 Associated Group, and 372 Field of Activity (these fields apply mostly to personal name and corporate name records). The level of application of a field 375 Gender, specific to personal name authority records, nearly doubled from 39% to 78%. The application of fields 374 Occupation and 368 Other Attributes increased minimally but remained substantial: 39.1% and 60.3% of records respectively. The level of use of 7 additional RDA-based fields remained low (under 1%). This is mostly explained by low proportion of uniform title and geographic name records in the dataset, as 6 fields are specific to representing works or places. However, given the high proportion of personal name authority records (39.6%), the level of application of 371 Address field is much lower than expected (0.065%).

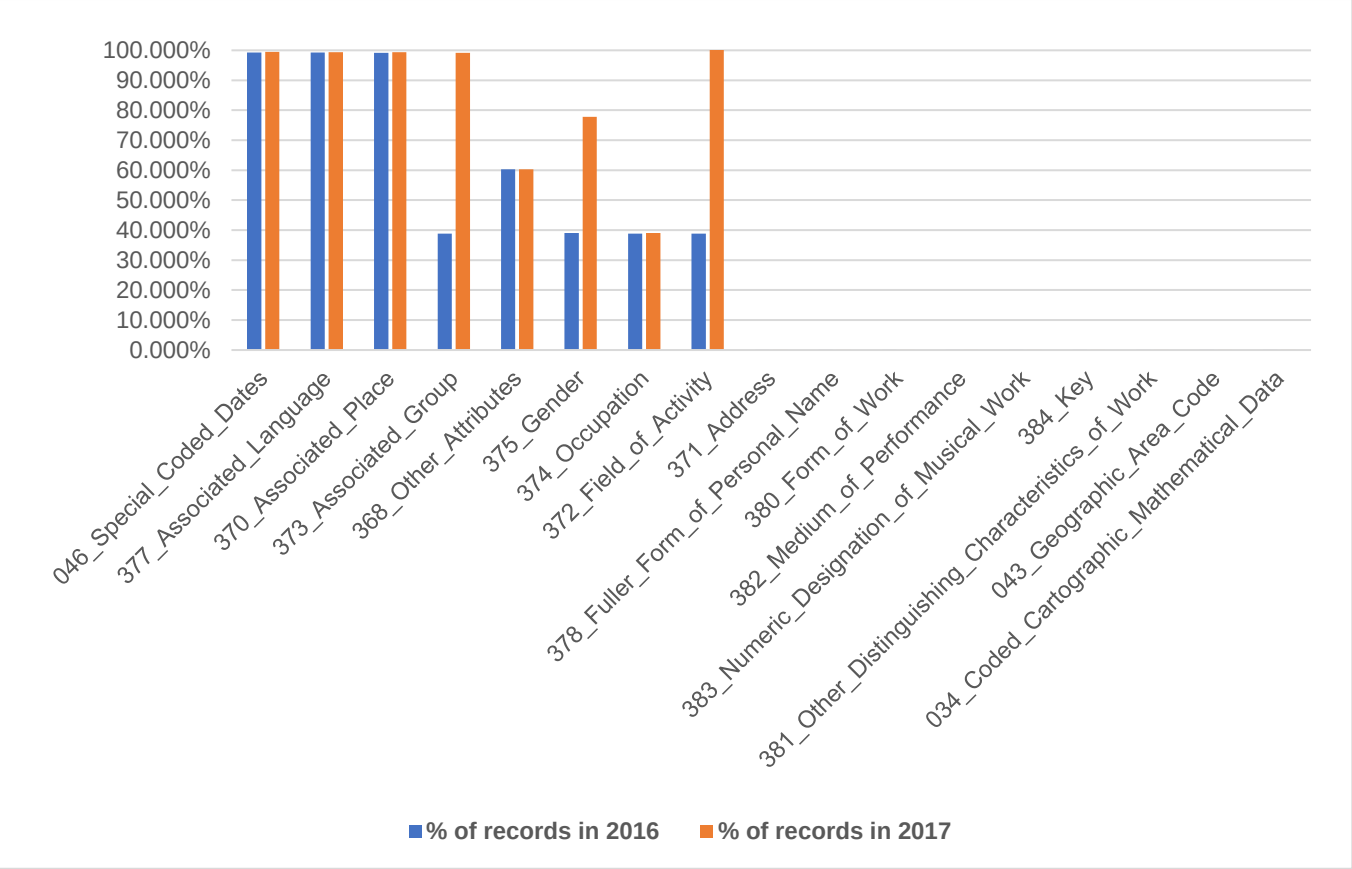


**Figure 1. Field-level data elements in RDA-based MARC fields (n=35472*)***

Over 60 subfields of RDA-based MARC fields occurred at least once in edited records. Some are repeatable so can occur more than once in a single instance of a MARC field in a record. Four of these data elements – mostly those applicable only to the uniform title records – were not present in 2016 versions of the records and have been added since: 371#v Source of [address] information, 382#b Soloist, 383#d Thematic Index Code, and 383#2 Source. Twenty data elements occurred significantly more frequently (in at least 38% of records) than others, the level of application of which was under 2%. Figure 2 shows distribution of top 20 RDA-based MARC authority subfields. The most noticeable increase in frequency of occurrence was observed for data elements that represent relationships: 370#c Associated Country, 372#a Field of activity, 373#a Associated Group, 374#a Occupation, and 375#a Gender. The use of the Linked-Data-enabling subfield #2 Source of Term in 3 RDA-based fields – 372, 373, and 375 the overall level of application of which has also significantly increased – has grown dramatically: tripled from 13.7K to 35.1K occurrences for 373, propelled from 2 to over 13.7K occurrences for 375 and from 89 to over 27.5K occurrences for 372. This Linked-data enabling subfield #2 was not observed in the generally infrequently-used field 383 in 2016 dataset but by late 2017, 18 occurrences were added to the records.

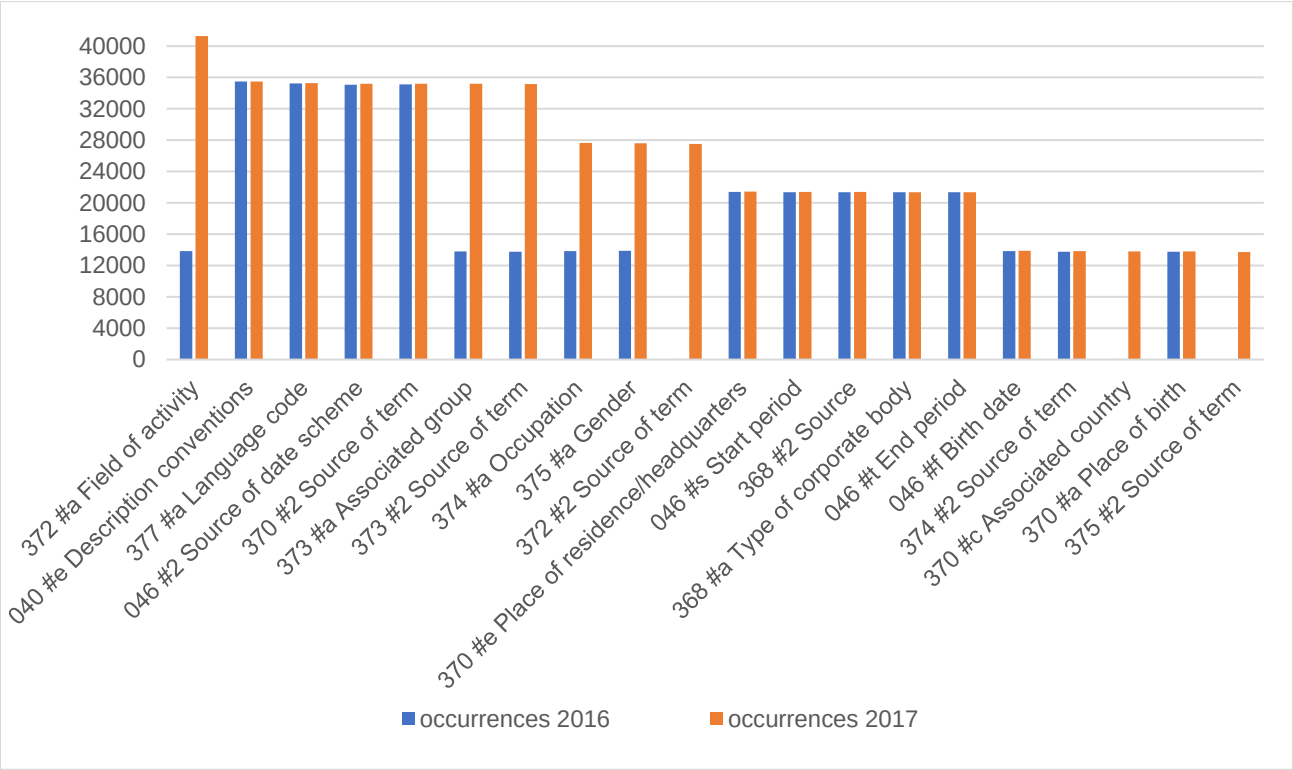


**Figure 2. Subfield-level data elements in RDA-specific MARC fields: top 20 (n=35472)**

RDA-based name authority records are also expected to contain new RDA-based subfields – mostly designed to support Linked Data functionality – in the traditionally-used pre-RDA fields (e.g., 410, 511, 670, etc.). This includes subfield #i for relationship information, subfield #u for unique resource identifier (URI), subfield #0 for authority record control number, subfield #1 for real world object URI, subfield #2 for information on the source of terms or other data, subfield #4 for URIs representing relationships. At the time of the first data collection in 2016, subfields #1, #4 and #i were not yet used to hold URIs and relationship information: #i and #4 were added to the MARC Authority standard in this function in May 2017, and #1 in December 2017. It is not surprising, therefore, that in late 2017 versions of authority records, no occurrences of Linked-Data-enabling elements #1 and #4 were observed. To the contrary, 4 pre-RDA fields (500, 510, 511, and 530) included subfield #i in high proportion (between 75% and 100%) of all records in the dataset containing these fields; for 3 of these fields, the level of use of #i increased between two data collection points. However, other Linked Data enabling RDA-based data elements, which existed in the standard for several years prior to our data collection, were found to be applied very infrequently in non-RDA-specific fields. For example, Subfield #u was observed in only one non-RDA based field: 670 Source Data Found but the level of its application increased by over 50%. Subfield #2 was found to be used, in a very small proportion of records, in only two non-RDA-specific MARC fields: 024 Other Standard Identifier (the level of application of 034#2 triples between two data collection points, and 034 Coded Cartographic Mathematical Data (the level of application remained unchanged)

## CONCLUSION

This study addresses the research gap in the area of implementation of RDA standard in authority data that are crucial for providing adequate access to information. Our quantitative analysis of change in RDA-based authority data reveals the lower overall level of editing activity than that observed by previous research for RDA-based bibliographic metadata (Zavalina, Zavalin, & Miksa, 2016). Results of our study demonstrate higher editing activity for meeting name and personal name authority data than for three other types of authority records. The change in application of certain data elements, related to evolution of RDA standard, was observed, with gradual and sometimes drastic increase in the use of elements representing persons, as well as some of the Linked Data-enabling elements. Despite the observed growth, the level of application of Linked-Data-enabling elements in authority records remains relatively low. More research is needed to longitudinally monitor the change in metadata records over time (especially those intended for providing Linked Data functionality). Comparative analysis of metadata change is also needed for different kinds of authority data and between authority data and bibliographic data. Qualitative analysis with the focus on data values in the record fields will supplement and help interpret quantitative results. These findings will help develop a solid understanding of metadata change and model it in relation to functionality of authority records and improved information access in digital libraries and beyond.

## REFERENCES


Kimura, M. (2015). A comparison of recorded authority data elements and the RDA framework in Chinese character cultures. *Cataloging and Classification Quarterly, 53*(7). DOI: 10.1080/01639374.2014.977984

Moulaison, H.L. (2015). The expansion of the personal name authority record under Resource Description and Access: Current status and quality considerations. *International Federation of Library Associations and Institutions, 41* (1), 13-24. DOI: 10.1177/0340035215570044

Niu, J. (2013). Evolving landscape in name authority control. *Cataloging and Classification Quarterly 51*(4), 404–419. DOI: 10.1080/01639374.2012.756843.

Southwick, S. B. (2015). A guide for transforming digital collections metadata into Linked Data using open source technologies. *Journal of Library Metadata, 15*(1), 1-35. DOI: 10.1080/19386389.2015.1007009

Stvilia, B., Gasser, L., Twidale, M., Shreeves, S., & Cole, T. (2004). Metadata quality for federated collections. In *Proceedings of the 9th International Conference on Information Quality* (*ICIQ04),* Cambridge, MA., *111-12*.

Stvilia, B., & Gasser, L. (2008). Value based metadata quality assessment. *Library & Information Science Research, 30*(1), 67-74. Retrieved from http://dx.doi.org/10.1016/j.lisr.2007.06.006.

Tarver, H., Zavalina, O.L., & Phillips, M. (2016). A case study of metadata creation in the University of North Texas Libraries' Digital Collections. *Proceedings of the 82nd International Federation of Library Associations World Library and Information Congress*, Columbus, Ohio, August 13-19, 2016.

Thompson, K. J. (2016). More than a name: A content analysis of name authority records for authors who self-identify as trans. *Library Resources and Technical Services, 60*(3). DOI: http://dx.doi.org/10.5860/lrts.60n3.140

Thornburg, G., & Oskins, M. (2007). Misinformation and bias in metadata processing: Matching in large databases. *Information Technology and Libraries, 26*(2), 15-26.

Zavalina, O.L., Kizhakkethil, P., Alemneh, D., Phillips, M., & Tarver, H.S. (2015). Building a framework of metadata change to support knowledge management. *Journal of Information and Knowledge Management, 14* (1), 1-16.

Zavalina, O.L., Phillips, M., & Tarver, H. (2017). Quality assurance and evaluation of change for patent metadata: poster abstract. *Proceedings of the 80th Association for Information Science and Technology Annual Meeting*.

Zavalina, O.L., Shakeri, S., & Kizhakkethil, P. (2015). Metadata change in traditional library collections and digital repositories: Exploratory comparative analysis. *Proceedings of the 78th Association for Information Science and Technology Annual Meeting*.

Zavalina, O.L., Shakeri, S., Kizhakkethil, P., & Phillips, M.E. (2018). Uncovering hidden insights for information management: Examination and modelling of change in digital collection metadata. In G. Chowdhury et al. (Eds.), iConference 2018, *Lecture Notes in Computer Science 10766* (pp.1-7). New York: Springer.

Zavalina, O.L., & Zavalin, V. (2017). Identity management analysis: an empirical investigation into the state of library community's' authority data conformance to the new standard. In D.G. Alemneh, J. Allen, & S. Hawamdeh (Eds.), *Knowledge Discovery and Data Design Innovation* (pp. 233-248). Hackensack, NJ: World Scientific.

Zavalina, O.L., Zavalin, V., & Miksa, S. D. (2016). Quality over time: A longitudinal quantitative analysis of metadata change in RDA-based MARC Bibliographic Records Representing Video Resources. *Proceedings of the 79th Association for Information Science and Technology Annual Meeting*.

Zavalina, O.L., Zavalin, V., Shakeri, S., & Kizhakkethil, P. (2016). Developing an empirically-based framework of metadata change and exploring relation between metadata change and metadata quality in MARC library metadata. *Procedia Computer Science, 99*, 50-63.